Motivated Metamodels

Paul K. Davis
Senior Scientist, RAND and Professor,
RAND Graduate School
1700 Main St. Santa Monica, CA 90407
(pdavis@rand.org)
Santa Monica, CA 90407
(pdavis@rand.org)

James H. Bigelow
Senior mathematician, RAND
1700 Main St., Santa Monica, CA
90407-2138
(James_Bigelow@rand.org)

ABSTRACT

A metamodel is a relatively small, simple model that approximate the "behavior" of a large, complex model. A common way to develop a metamodel is to generate "data" from a number of large-model runs and to then use off-the-shelf statistical methods without attempting to understand the model's internal workings. It is much preferable, in some problems, to improve the quality of such metamodels by using various types of phenomenological knowledge. The benefits are sometimes mathematically subtle, but strategically important, as when one is dealing with a system that could fail if any of several critical components fail. Naïve metamodels may fail to reflect the individual criticality of such components and may therefore be misleading if used for policy analysis. By inserting an appropriate dose of theory, however, such problems can be greatly mitigated. Our work is intended to be a contribution to the emerging understanding of multiresolution, multiperspective modeling.

Keywords: *metamodel, multiresolution modeling, model abstraction, response surfaces, repro models, statistics, regression, intelligent machines, robotics, machine planning*

INTRODUCTION

A metamodel is a relatively small, simple model intended to mimic the "behavior" of a large, complex model. Two reasons for wanting to build metamodels are.[1]:

- Cognitive. We want to "understand" why the large model behaves as it does. This will enhance the model's meaningfulness and credibility with ourselves, other analysts, and with whomever we seek to influence.
- Exploratory analysis. We often want to explore the behavior of a model over a large part of its domain. A metamodel with only, e.g., 5-10 (rather than hundreds or thousands) of variables makes comprehensive exploratory analysis feasible and comprehensible[2], [3].

Sometimes, it is possible to build models using multiresolution modeling design principles [1], in which case the low-resolution versions (the more abstracted or aggregate versions) already have these virtues. Often, however, the baseline situation is that a subject area has been modeled only in detail. Often, the detailed model is old, opaque, and difficult to work with. A metamodel, then, is an attempt to generate a low-resolution version after the fact.

Consider first two extreme approaches to building a metamodel. For the *statistical* approach, one runs the large model many times for a variety of inputs; one then collects the inputs and outputs in a big dataset. A statistician analyzes these data as he would data collected in any series of experiments. He seeks a statistical model that does a good job of estimating outputs, while using as few input variables as possible so as to keep the model simple and so as to avoid "overfitting" his initial data.

An idealized *phenomenological* (or theory-driven) approach starts with the most exact theory available and derives a simplified model by rigorously aggregating (as in replacing integrals over volume with a representative value times the volume), rearranging and combining terms and factors, and so on—using physical insights wherever possible (e.g., recognition of some conservation principles or of being able to view a factor as an idealization times an efficiency factor).

Some statisticians, operations researchers, and computer scientists prefer the first approach and want to know nothing about the "innards" of the model whose behavior they are attempting to replicate. They may have a purist philosophy of "allowing the data to speak," without "contaminating it" with theoretical assumptions. Or they may simply prefer not having to deal with the complexities of the model's innards: they may wish to turn the problem over to automated software. At the other extreme, some theoretically inclined academicians clearly prefer the second approach because it allows rigorous tying together of phenomena at different levels of detail (as when classical thermodynamics is understood from quantum mechanics). These, then, are the extremes. Most scientists, engineers, and analysts, however, should prefer something in between. Often, the "in-between" amounts to an analyst postulating some simple scaling relation and using data to calibrate that scaling relation. The scaling relation may be naïve and the calibration may use only the crudest of statistics.

Our interest in this research was to clarify principles for doing better. We had some ideas for how to do so, based on theoretical reasoning, but we preferred where possible to test and iterate ideas by experimenting with specific well-posed problems. With that in mind, we used a well-documented model that we could consider "large and complex model" and applied our ideas in stages, starting with pure statistical metamodeling and moving toward more theory-informed work. In what follows, we first discuss how one should judge the quality of a metamodel. We then describe our experiments and conclusions. The discussion adds to an earlier preliminary discussion [4] and is drawn from a longer and more technical piece [5].

WHAT MAKES A METAMODEL GOOD

We suggest five criteria for assessing the goodness of a metamodel, heroically assuming for the sake of this discussion only that the base model is in fact accurate:

Goodness of fit. Obviously, we want the metamodel's predictions to be reasonably consistent with those of the baseline model. A measure of this is the root mean square error (or fractional error) of predictions across the domain of input values. This is superior to the commonly used R^2.

Parsimony. For purposes of both cognition and exploratory analysis, a good metamodel will have relatively few independent variables, which ideally, would also be meaningful. Achieving parsimony may be accomplished by omitting some of the baseline model's inputs (i.e., treating them as constant) or by combining several base model inputs into a smaller number of intermedate variables. The set of independent variables should be rich enough to represent the issues being addressed with the model. Beyond that, the fewer extra variables, the better.

Identification of "critical components." Our third criterion seems new and we believe it to be crucial. Many uses of models in analysis involve systems or strategies, the failure of which is to be very much avoided. We suggest that a metamodel should highlight all of the input variables that are essential to success—especially when troublesome values of those variables are plausible. The model should not give the impression that one can compensate for a weak component of the system by improving some other component (if such substitution is in fact inadequate). This is a significant consideration in metamodeling, because standard statistical methods lead to linear sums that imply substitutability. We refer to components that must individually succeed (have values above or below an appropriate threshold) as *critical* components. If critical components in this sense exist, the metamodel should be appropriately nonlinear.

Reasonable depiction of relative "importances." Metamodeling can generate statistical measures of the significance or importance of candidate variables. In stepwise regression, the less significant variables are dropped. That, in turn, could mean that someone using the metamodel for resource allocation would consider the dropped variables as unimportant. A good metamodel would give no misimpressions on this score.

A good storyline. Without a story, a model is just a "black box." A story explains *why* the model behaves as it does. More, it relates the model to the real world, telling us why the model *should* behave as it does. We use the term "storyline" because all models are a simplification of reality, but we intend no cynicism. Said differently, the model should be "physically (or otherwise phenomenologically) meaningful and interpretable," not just a math formula.

With this background defining what we mean by a good metamodel, let us now describe the analytical experiments we conducted to illustrate and sharpen our understanding of ways to improve metamodeling.

THE EXPERIMENT

Our experiment was to begin was a relatively large and complex model and to develop a series of metamodels to represent it. For the first metamodel we relied almost entirely on statistical methods, uninformed by phenomenology (i.e., our knowledge of the workings of the base model). With each successive metamodel, we took advantage of progressively more phenomenology.

The Large Model

Our "large, complex model" (i.e., our baseline model) was EXHALT-CF [5],[6], which treats the so-called halt phase of a military operation. Although much simpler than real base models of interest, it has scores of variables and a great many nonlinearities. It seemed complex from its documentation and program.

In its simplest version, the halt phase is a mere race. An attacking force (Red) is advancing on an objective while the defenders (Blue) interdict its armored vehicles with long-range fires. Red will halt when he reaches his objective (a Red win) or when Blue has killed a specified number of vehicles (a Blue win), whichever comes first. EXHALT-CF, however, adds many embellishments relevant to current strategic concerns about real-world military operations, especially in the Persian Gulf.

First, the model must represent Blue deployments. Some number of shooters may be stationed in theater in peacetime. Depending on strategic warning, diplomatic relations, Red's deceptiveness, and Red's ability to threaten bases in theater (e.g., with weapons of mass destruction), Blue may or may not be able to augment this number before Red begins his advance. Once Red's advance begins, Blue will deploy more shooters into the theater, up to a theater capacity, which reflects logistical shortcomings.

The effectiveness of Blue shooters is measured by kills per shooter-day. Early in the campaign, Blue may be unable or unwilling to attack the Red column because of Red air defenses. After a period of air-defense suppression, Blue's attacks will start. Even then, however, sortie rates may be reduced because of a continued threat of attack with mass-destruction weapons, which would force Blue personnel to work in protective gear or would force Blue to operate from more distant, and more poorly prepared bases.

The weapons and strategy Blue selects will also influence Blue shooter effectiveness. Blue may select an area weapon, capable of killing several Red armored vehicles per shot. To counter this, Red may space his vehicles more widely. Or Blue may select a point weapon, which kills no more than one vehicle per shot, and is unaffected by Red's vehicle spacing. Also, Blue will likely have limited supplies of his best weapons, and revert to lesser weapons when his best are exhausted. Blue may attack the entire Red column in depth (the "In Depth" strategy) or focus his attack on the leading edge (the "Leading Edge" strategy). If Blue does the latter, his attack may slow Red, but each sortie may be less effective due to deconfliction problems.

These and other complications of the halt problem are represented in EXHALT-CF and the simulation version, EXHALT. They are implemented in Analytica□ , a graphical modeling environment for the personal computer. EXHALT-CF has 63 inputs, 8 switches to turn features on or off (the model has a multiresolution, multiperspective design) , three indexes, and 451 variables that are calculated directly or indirectly from inputs. For our purposes, we focused on a subset of cases, which reduced to 25 the number of input variables affecting the problem. This seemed adequately complex to illustrate our points—or, more accurately, to allow us to experiment. The experiments in question were experiments of discovery and learning, not rigorous hypothesis testing.

The Experimental Data

We selected statistical distributions, mostly uniform distributions, from which to generate the 25 variable-value inputs. We then ran EXHALT-CF to generate a Monte Carlo sample of 1000 cases from the overall input space. We did not weight one or another region of the input space because we were seeking a broadly good fit of behavior over the entire domain of interest.

METAMODELING

Metamodel 1

In our first experiment, we acted as though we had handed the dataset to a statistician (or statistically oriented operations researcher or computer scientist), and commissioned him to develop an estimator for the halt distance that Blue could achieve using his best strategy and weapon type for the circumstances of the case. A good statistician would insist on discussing the problem. He would want to know which data elements to use as independent variables, which are outcome variables, and so on. Even if he preferred to operate as thought the original model is a "black box," he would probably want at least some interpretation of the variables' meanings. This would permit him to do some data manipulation that would simplify his analysis.

For his initial analysis, our simulated statistician specified a linear model with 25 independent variables. Because he knew that we wanted a parsimonious metamodel, he ran a *stepwise linear regression* procedure in which the independent variables were added to the model one by one in the order of decreasing explanatory power. That is, the first variable considered yielded the largest reduction of the root mean square of the residual error (RMS Error. After the first six or seven variables, further additions didn't improve the fit very much. Actually, the fit wasn't very good no matter how many variables were included (the standard error was on the order of 25%, which in this problem is large).

In any case, our simulated statistician stopped with 14 variables, all coefficients of which were significant at the 0.05 level.

How good was metamodel 1? Earlier we identified five features that make a metamodel good. The performance was not impressive, although, in our experience with metamodeling pure statistical approaches such as this not uncommonly do quite well by the average goodness-of-fit criterion. Still, in this case, there were 421 cases in which Red actually reached his objective and the model estimated that Red was halted short of his objective in 92 percent of them.

Parsimony was our second criterion for a good metamodel. This model had 14 variables. We would like fewer, but this was perhaps a marginally acceptable number.

Identification of critical components was our third criterion. Here Metamodel 1 performed very poorly. Such metamodels are linear in the variables identified as significant by the statistical analysis. Thus, when used, metamodel 1 failed to identify and highlight critical components. For example, the model would predict that by merely improving munitions sufficiently, Blue could guarantee a small halt distance—independent of the other variables. That is flatly wrong. The model also gave a very misleading sense of the relative importance of variables. After stepwise regression, one variable had dropped out, while another—which entered the problem in precisely the same way (if the variables were X and Y, they entered the problem only through the product XY)—was retained as significant. This could be a very serious shortcoming if the model were used to inform resource-allocation decisions. Also, as discussed earlier, ur final criterion for a good metamodel was that it have a good storyline. This metamodel had no storyline at all.

Metamodel 2

A statistician will often try to improve his model by introducing transformations of the independent variables, such as exponentials, powers, and products of variables. That is, he will still use linear regression, but with some of the variables of that regression being nonlinear composites of others. So many possible transformations of variables are possible that the statistician may need some guidance selecting which ones to try. Brute force (e.g., considering all of the quadratic combinations of elementary variables as new, composite variables) can result in a good fit, but usually with even more statistically significant variables and no "story."

Phenomenology (i.e., the "innards" of the baseline model) can suggest what transforms to try, including transforms that statisticians do not generally consider. These include transforms that use the MAX and MIN operators. Indeed, a number of transformations are built into EXHALT-CF. We designed it as a *multi-resolution* model, to permit the user to specify inputs at different levels of detail. Even if the metamodeler finds EXHALT-CF as a whole to be big and complex, even early chapters of documentation, which deal with various idealizations, are sufficient to highlight natural composite variables. They may not *fully* substitute for the more elementary variables, because the "real" EXHALT-CF (as distinct from the simplified versions discussed in early documentation chapters) includes more complex interactions. Nonetheless, we thought that the suggested composite variables ("aggregation fragments") might go a long way.

For Metamodel 2, then, we looked at a number of such composites."

The simulated statistican then defined a linear model with far fewer variables, many of them the composites.

How good was metamodel 2? The performance, while considerably better than Model 1, was still not impressive. The standard error was perhaps 80 km, rather than 140 km (with interesting values being in the range 0-600).

On grounds of parsimony the model was better, since only ten independent variables proved to be statistically significant.

The good news was that the model did predict the critical-component phenomenon: we had identified enough of the key composite variables so that we could see importaant nonliearities. In particular, to obtain a good halt distance, Blue had to address three issues simultaneously (with no substitution). These involved the number of initial shooters, the earliest time at which Blue could begin attacking Red effectively, and the number of "shooter days" required for success (a functon of Red's size and Blue's effectiveness per attack mission). This important "system feature" stood out. However, this metamodel still did not have a storyline, although the variables at least had more physical significance.

Metamodel 3

So far the simulated statistician had been combining the original, low-level variables into intermediate variables that we thought were reasonable on phenomenological grounds. We might characterize this as a "bottom-up" strategy. Now we turned to a "top-down" strategy. We viewed this as explicitly building in a storyline. It depended on an understanding of phenomenology, but it not require that the theory for describing that phenomenology be analytically tractable (e.g., solvable in closed form).

One piece of knowledge used was inferrable from even minimal documentation of EXHALT-CF, to wit that the model considered two different Blue strategies and took the better of the two results as the answer. The model did similarly in comparing two different classes of weapons. In the earlier metamodels, these branches and use of MIN/MAX operators was all buried, but in Metamodel 3 we built that same macro logic in. This meant that Metamodel 3 actually involved four metamodels, plus logic to compare results from each.

More important, we inserted physical reasoning in simplified terms. Upon thinking about the problem physically, we could reason that the halt time was just the time required to kill the requisite number of Red targets. However, that depended on the number of Blue shooters, which increased linearly with time (subject to some further constraints in the full model), the size of the Red force, and the per-shooter-day effectiveness of Blue. We could write a simple analytical equation for this—so long as we glossed over details and inserted averages. We estimated an integral by the product of a time duration and the average number of shooters within the interval (without knowing precisely how to calculate that average).

We then made a very crude estimate of this and other averages. The result was dimensionally correct and not absurd, but it was not intended to be accurate.

What we were looking for was *structural* form. This we postulated as the basis for Metamodel 3, although building in fudge factors to measure error in the form assumed.

How good was metamodel 3? Metamodel 3 fitted the data much better than either of the two previous metamodels. It was also much more parsimonious than the previous metamodels. It had only five significant variables and the fudge factors proved to be not very large. Thus, the "story" understandable from the highly simplified model that did violence to mathematics by, e.g., treating an integral as a product of a duration and an average value during the period, was largely correct.

The model also did well in predicting the critical-component phenomenon. The composite variables that made ths possible in Metamodel 2 were also present in Metamodel 3, but now with a better story. Nor were there any serious errors with respect to the relative importance of variables. All in all, results were rather good: building in the approximate structure had paid off handsomely.

Metamodel 4

The last metamodel that we considered pushed the analytic work further. Upon reflection, it was possible—in this particular problem—to do a much better job of estimating the average number of shooters present during the halt phase. This required nothing more profound than simple integration and relatively simple algebra. A good student of first year calculus would be able to do the problem without difficulty. If we inserted this knowledge, the resulting metamodel was exceedingly accurate—so much so that it was embarrassing. The results demonstrated that the complexity of EXHALT-CF was the result of essentials having been obfuscated by the programming. The subtleties and complications were simply not necessary when the numbers were crunched.

Interestingly, however, Metamodel 4 was *not* better than its predecessor by our criteria. Why? Because, in the process of inserting the improved solution, the structural form became more complicated and non-transparent. This obscured the story, making it impossible to actually put the model up on viewgraph and explain what was going on as, e.g., a product of three meaningful variables divided by a fourth, with a small error term reflecting the many simplifications involved. If we wanted clarity and insight, then Metamodel 3 was arguably better. We say "arguable," because with clever presentation one could hide some of Metamodel 4's complications.

SUMMARY AND LESSONS LEARNED

Our experiments confirmed our belief that much could be gained by combining virtues of statistical and phenomenological (theory-informed) approaches. They confirmed and give more precise arguments to our skepticism about approaching metamodeling as an exercise in pure data analysis, with the baseline model merely being a black-box generator of data. Although the experiment dealt with only a single baseline model, the insights appeared to us to be relatively general—at least for purposes of suggesting general cautions and approaches to consider. We intend no grand claims here, but those were (and are) our impressions.

In our experiments, the application of statistical methods uninformed by phenomenology did not produce a good metamodel. In part it failed because the data we were fitting describe a highly nonlinear surface. A linear model might fit locally, but never globally. Moreover, there was no guarantee that the regression coefficients would be a good guide to the relative importance of the independent variables. It was necessary to introduce nonlinear combinations of the low-level inputs in order to obtain a good fitting metamodel. Phenomenology could motivate the construction and selection of the appropriate nonlinear combinations. It would be very difficult to discover them from the data alone.

A linear model also failed because the data did not describe a smooth surface. Like many models, EXHALT-CF makes liberal use of MAX and MIN operators to make either-or choices. So the data described a surface with "kinks" in it. When one fits a kinky surface with regression, the coefficients obtained from regression don't need to make sense. The regression results tell us the average importance of inputs, but not *when* each one is important. We found it necessary to use phenomenology to separate the smooth segments of the surface, after which we could do a good job of fitting each smooth segment with just plain statistics.

As we introduced more and more phenomenology, we obtained successively better fitting models. We would argue that in addition, Model 1 had very little cognitive value (i.e., it didn't tell a story) and Model 2 had only a little more. Model 3, however, did tell a coherent story, one that could be related persuasively to the client. With Model 4 it could be argued that we began to lose cognitive value. The phenomenology became more complex and less transparent. All the equations began to obscure the story.

Particularly important also is that by inserting phenomenologically motivated structure one can avoid certain important blunders of system depiction. In particular, one can preserve and even highlight the role of critical components—components that enter the problem more nearly as products than as sums, or components that must individually have threshold values to avoid system failure. This is particularly important if metamodels are to be used in policy analysis or design. We would expect it to be quite important in design of intelligent systems, because we would expect it to be normal, not unusual, for designers to be worried about numerous independently critical factors.

ACKNOWLEDGMENTS

This research was sponsored by the United State Air Force Research Laboratory.

REFERENCES

[1] Davis, P. K. and J. H. Bigelow, *Experiments in Multiresolution Modeling*, RAND, Santa Monica, CA, 1998.

[2] Davis, P.K., J. H. Bigelow, and J. McEver, Exploratory Analysis and a Case History of Multiresolution, Multiperspective Modeling, RAND RP-925, Santa Monica, CA. Reprinted from *Proceedings of the 2000 Winter Simulation Conference*, J. A. Joines, R. R. Barton, K. Kang, and P.A.. Fishwick (editors), December, 2000 *and Proceedings of the SPIE*, Vol. 4026, 2000.

[3] Davis, P. K., J. McEver, and B Wilson, *Measuring Interdiction Capabilities in the Presence of Anti-Access*

Strategies: Exploratory Analysis to Inform Adaptive Strategy for the Persian Gulf, RAND, Santa Monica, CA MR-1471-AF, 2002.

[4] Davis, P.K. and J. H. Bigelow, "Meta Models to Aid Planning of Intelligent Machines," Proceedings of the 2001 PerMIS Workshop, September 4, 2001 (NIST Special Publication 982), Gaithersberg, MD.

[5] Bigelow, J.H. and P.K. Davis, "Developing improved metamodels by combining phenomenological reasoning with statistical methods," *Proceedings of the SPIE*, Vol. 4716, 2002 (A. Sisti and D A. Trevasani, editors).

[6] Davis P.K, J.H. Bigelow, and J McEver, *EXHALT: An Interdiction Model for Exploring Halt Capabilities in a Large Scenario Space*," The RAND Corporation, MR-1137-OSD, 2000